COMPARING THE THINK ALOUD INTERVIEWING TECHNIQUE WITH STANDARD INTERVIEWING IN THE REDESIGN OF A DIETARY RECALL QUESTIONNAIRE

Wendy L. Davis and Theresa J. DeMaio
Wendy Davis, U.S. Bureau of the Census, CSMR, WPB1, Room 433, Washington, DC 20233-4700

Cognitive interviewing techniques are quickly gaining recognition as useful methods for pretesting and designing questionnaires. The National Center for Health Statistics (NCHS), the Bureau of Labor Statistics (BLS), and the Bureau of the Census have all set up cognitive laboratories for pretesting and designing questionnaires. Conducting cognitive interviews to test a questionnaire allows researchers to understand how respondents interpret a question and what mental processes they must engage in to answer the question. With this information, researchers can assess whether questions are being interpreted in the intended manner, whether respondents are able to answer certain questions depending on its cognitive demands, and so on. Some examples of work done using cognitive interviews for testing questionnaires are Jobe and Mingay (1989, 1990), and Willis, Royston, and Bercini (1991)[1]. Depending on the time and funding available, cognitive interviews may either be used alone or as a predecessor to field tests or other pretesting methods. The advantage of using other pretesting methods such as field tests is that they allow the researcher to see how the questionnaire works in a standard interview format outside of a controlled laboratory setting. In addition, after conducting cognitive interviews and making changes to the questionnaire based on these interviews, a field test provides some feedback on recommended changes.

Field tests are usually conducted either with a single, newly-revised questionnaire or as a split-panel experiment in which the revised questionnaire is tested against the original version. However, there is little work to date that systematically investigates whether results from cognitive interviewing generalize to a field setting. This is particularly important for surveys which, due to time constraints, are only pretested in a laboratory setting. It could be that the observations made in a controlled laboratory environment using cognitive interviewing techniques are not applicable to a field setting since the response process of the two interviewing methods are very different. This paper compares cognitive interviewing techniques with standard field interviewing techniques using a dietary intake questionnaire.

The remainder of the paper presents some background information addressing the initial cognitive research done using the dietary intake questionnaire, some discussion of cognitive versus standard interviews, and a description of the present research comparing the two interviewing methodologies. Finally, the implications of adopting recommendations from cognitive interviews without first testing them in a field setting are discussed.

Cognitive Research Using the CSFII Instrument

The Continuing Survey of Food Intakes by Individuals (CSFII) is a national survey conducted once every three years by the U.S. Department of Agriculture Human Nutrition Information Service (HNIS). It is designed to collect the dietary intake of individuals over a twenty-four hour period as well as gather some other general health and nutrition information. HNIS asked the Center for Survey Methods Research, U.S. Bureau of the Census, to conduct cognitive research on the instrument and to make subsequent recommendations about how to improve response accuracy and minimize respondent burden. After reviewing the CSFII questionnaire, we decided that by focusing our research efforts on the dietary intake portion of the instrument, we could address both of these issues.

The dietary intake portion of the questionnaire is primarily a recall task. Among other things, the respondent is asked to remember and report, in great detail, all of the foods and beverages consumed in a 24-hour period, the time each item was consumed, the amount consumed, and the place where the food or beverage was obtained (e.g. grocery store, restaurant). To answer these questions, the respondents must search their memories for the target information. Previous research has shown that memory retrieval can be

enhanced or diminished, depending on the types of cues provided to the respondent. Means et al. (1991) and Tulving (1983) did studies which suggest that incorporating contextual information from the target event aids in the recall of specific information concerning that event. For example, in the Means et al. study of smoking cessation behavior, they found that first asking respondents about their reasons for quitting smoking and about the family support they received improved their recall of the specific date when they initially tried to stop smoking. In line with this work, it seemed that in order to improve response accuracy, we needed to get respondents to think about the relevant contextual information surrounding each time they consumed a food or drink. However, there was no information available to us that suggested what the relevant contextual information would be. So our first goal was to collect data about what information people naturally use to help them remember their food intake. From there we would redesign the questionnaire to include this information as part of the questions themselves.

We conducted 6 think aloud interviews that allowed the respondents to freely recall the foods and beverages they consumed the previous day using the memory retrieval cues that came most naturally to them. In other words, the respondents could think about anything they wanted to help them remember what they had consumed. Our initial idea was to incorporate the most commonly used recall strategy into the questionnaire. However, after these interviews it became obvious that across respondents there really wasn't a "most common" recall strategy. This being the case, incorporating any one strategy into the questionnaire may actually hinder the recall for some respondents who would not naturally remember their intake using that retrieval strategy. So, instead we decided to write a question that let respondents choose their own method of recalling the foods and beverages consumed the previous day and report these foods in any manner they wanted. The question read "Tell me everything you had to eat or drink yesterday, from midnight to midnight. Include everything eaten at home or away--even snacks, coffee breaks or alcoholic beverages." After testing, however, we realized that this wording may suggest a chronological review of the day by specifying "from midnight to midnight." It also may suggest that the respondent should associate the day's activities with the foods they consumed by indicating they should include everything eaten "at home or away." The implications of this wording are discussed below.

We conducted 11 more think aloud interviews using the free recall question as the first question in the interview. After completing 17 cognitive interviews in total, we recommended improvements to be made to the questionnaire. The recommendations consisted of specific wording changes, as well as recommendations for changing the order of questions in a way that seemed to compliment the respondent's cognitive processes. In order to meet the operational demands of the survey, however, we were not able to conduct any further research to evaluate our recommendations.

Cognitive Interviews vs. Standard Interviews

Cognitive and standard interviewing techniques differ in many ways, but for the purposes of this paper, we will focus only on three differences between the two techniques.

The first difference is in what is emphasized to the respondent as important. In a concurrent think aloud interview with probing, respondents are told that what is of the most interest to the researcher is the cognitive process that they go through as they form an answer, rather than the answer itself. Typically, respondents are told this at the beginning of the interview as a general introduction to the task, and again during the interview as an introduction to specific questions, or as feedback throughout the interview in the form of probing questions. Cannell et al. (1981) found that when interviewers gave an introduction before questions and provided feedback, the accuracy of responses was improved. A standard field interview, on the other hand, typically does not include any probing questions, and often does not have any introduction prior to individual questions. When in the field the dietary intake instrument in the study described in this paper includes neither introductions before questions nor feedback.

The second way in which the two techniques differ is in the overall pace of the interview. A cognitive interview is typically longer than a standard field interview. In fact, at the Bureau of

the Census, we expect a cognitive interview to take twice the amount of time as a standard interview. Burton and Blair (1991), though not using cognitive interviews, and Means et al. (1989) found that the more time respondents were given to answer, the more accurate their answers.

The third difference between the two techniques is in the amount and type of information retrieved from memory. In a concurrent think aloud interview, respondents are told to report aloud everything that they are thinking, which in turn slows down the pace of the interview. Focusing on your cognitive processes enough to report them aloud, and as a result, slowing down the response process may cause certain information to become more salient than it would had the respondent not been thinking aloud. For example, when people first learn how to play golf, their instructor may have them practice their swing at an exaggeratedly slow pace so that they will notice when they shift their weight or when their eyes leave the ball. The same may be true when thinking aloud. Certain information in memory may become more salient than it would during a quicker, less intensive standard interview. Kahneman and Tversky (1971) found that the information most salient in memory is also more available and thus more likely to be recalled. On the other hand, some researchers have postulated that during a field interview, respondents may actually do the minimum amount of processing possible in order to come up with an answer. Alwin and Krosnick (1987) refer to this as being a "cognitive miser."

Since our recommendations for changes were based only on data collected using cognitive think aloud interviews, we were concerned that some of the success we had in getting detailed reports from respondents might have been due to the think aloud process itself rather than the instrument. The objective of this study was to investigate whether think aloud interviews elicited reports of more foods and beverages consumed the previous day than standard interviews.

Research Study

Procedure

The ideal way to test whether thinking aloud improved recall would be to compare the reported recalls given under instructions to think aloud and standard interview conditions and then validate the reports. However, a validation study is beyond the technical means of such a small scale study. So instead, we relied on portions of the redesigned instrument to investigate the effects of thinking aloud.

To do this we divided the questionnaire into two parts, the first question being the first part, and questions 2 through 9 making up the second part. Item 1, referred to as the Quick List, asks respondents to report, in any manner they choose, all of the foods and beverages consumed the previous day. The question reads "Tell me everything you had to eat or drink yesterday, from midnight to midnight. Include everything eaten at home or away - even snacks, coffee breaks, or alcoholic beverages." The Quick List allows respondents to do a free recall and allows the researcher to find out what and how many items the respondents can remember without any additional memory cues.

Items 2 through 9 then follow up on the foods and beverages reported in the Quick List with specific questions, such as the time the food item was consumed, the name of the eating occasion (e.g. breakfast, supper, snack), where the food item was obtained, etcetera. Each of these questions directs the respondents' attention to something specific about the occasion when the respondent consumed a food item. Since these questions give respondents another opportunity to review the day's intake and report any items that may have been forgotten, the end result is an additional list of foods including and expanding on the original list given in the first question. Work by Laurent, Cannell and Marquis (1972), Means and Loftus (1991), and Means et al. (1991) supports the use of multiple questions to elicit information about a single event in order to improve recall of that event. In other words, the purpose of the Quick List was to get the initial recall of the previous day's intake, and items 2 through 9 were to get all the remaining items that were forgotten in the Quick List. In the actual administration of the survey, this combination of methods was designed to elicit as accurate a report of food intake as possible (see DeMaio, Ciochetto and Davis, 1993, for further description of the rationale and content of the questionnaire revisions). For the purpose of this research, however, we are using questions 2 through 9 to evaluate the information

provided in question 1, since we expect question 1 to be the most affected by the cognitive interviewing procedures, given its less directive nature.

With the instrument divided in this manner, we then split our respondents into two groups of ten each. The first group, the experimental group, was instructed to think aloud, and the other group, the control group, was simply given the standard introduction to the study. Each group was asked the first question and allowed as much time as they wanted to respond. Then, before proceeding to the next portion of the interview, the interview was stopped for 10 minutes, during which the respondents were given a separate questionnaire about general health and nutrition to complete on their own. The questions contained in this additional questionnaire are included on the actual CSFII questionnaire and asked after the 24-hour recall has been administered. The point of this interruption was to distract the experimental group so that when answering the focused questions, items 2 through 9, they would be thinking in the same manner as the control group and no longer thinking aloud. The time it took to complete the supplemental questionnaire varied from 4 to 10 minutes, though the interviewer always waited 10 minutes before returning to the room and resuming the interview. The interviewer began the second portion of the interview without mentioning thinking aloud. In fact, not one of the respondents in the experimental group questioned whether they should continue to think aloud. All respondents completed the second portion of the questionnaire in a standard interview format.

After completing twenty interviews in total, each interview was coded by both authors independently and differences were resolved. Each interview was coded for the number and type of foods, beverages and eating occasions that were reported in the Quick List and in questions 2 through 9. Age, gender, race and whether or not the respondent was on a diet were also recorded.

Results and Discussion

Given the small sample size, our study is exploratory in nature. We have included statistical tests of observed numerical differences, none of which are statistically significant. Thus, our results and conclusions are suggestive rather than definitive.

There were a total of twenty respondents, all of whom were paid volunteers recruited from advertisements in local newspapers. Table 1 contains their demographic characteristics. Overall, there were more female respondents than males. The groups were equally divided between white and non-white respondents. As the table shows, the characteristics of respondents in the two treatment groups were similar.

Table 1: Demographic Characteristics by Group

	Standard	Cognitive	Total
Mean Age	38.6	37.2	
Male	3	1	4
Female	7	9	16
White	7	4	11
Non-white	3	6	9

There are two main comparisons to examine our hypothesis that thinking aloud may enhance recall of foods and beverages. The first compares the mean number of items reported by each group in response to the first question. We expected the cognitive group to report a larger number of items than the standard group. Secondly, the mean number of ALL foods and beverages reported by the end of the interview should be equal across the two groups, since the second part of the interview was designed to capture any foods forgotten in the first question. Thus, the only expected difference between the two groups was in response to the first question.

We did not find a difference in the predicted direction. As Table 2 shows, the mean number of items reported in the Quick List for the standard interview group was 11.4 versus 10.4 for the cognitive group, a nonsignificant difference.

Table 2: Mean Number of Items by Introduction Type

	Standard	Think Aloud	T	$p > \|T\|$
Quick List (s.d.)	11.4 (2.5)	10.4 (3.2)	0.77	0.45
Final List (s.d)	19.5 (6.8)	24.6 (7.5)	-1.59	0.13

As part of the design of our experiment, we are assuming that the second portion of the interview will elicit a complete recall of the previous day's dietary intake, so the total number of foods and beverages reported by the two groups should be equal by the end of the interview. Once the experimental manipulation (thinking aloud) is finished, the second part of the survey should affect the respondents of both groups in the same manner. Table 2 reveals a nonsignificant difference, a mean of 24.6 foods and beverages for the cognitive group and 19.5 for the standard group.

If this numerical difference were an indicator of a true difference, the direction of the results is surprising since it is in the opposite direction from the results observed with the Quick List. Although at this point in the interview both groups were proceeding in a standard interview format, it is possible that there was a delayed reaction among respondents who initially received the instructions to think aloud. While none of the respondents continued to think aloud once the interview resumed after the ten minute break, they may have taken the task more seriously and searched their memories more carefully after the initial think aloud instruction.

An alternative explanation for the numerical difference does not involve the recall task but rather lies in the rules for reporting food items, specifically home-made food items. To maximize the amount of detail reported for a food item, at question 4 (in the second part of the interview when all respondents are reporting in a standard interview format), respondents are requested to report all the ingredients of a food or beverage if they made the food or beverage themselves. Those who did not prepare the food or beverage are requested to report only the single item. For example, if two respondents report having stir-fry chicken, but only one of the respondents prepared it, the number of food items recorded for each respondent would be different. The respondent who prepared the stir-fry him/herself would be specifically asked to report all of the ingredients such as chicken, pea pods, water chestnuts, broccoli, etc., whereas the respondent who ate the same meal at a restaurant would simply report chicken stir-fry. In the final count of food items then, the respondent who prepared the chicken stir-fry would be credited for reporting more items than the respondent who did not prepare the dish. This is a misleading number in terms of recall, since both respondents were able to remember and report that they had chicken stir-fry the previous day.

To control for this difference in "countable" items, we eliminated all but one of the ingredients that were part of a recipe. The results of this analysis are shown in Table 3 below.

Table 3: Mean Number of Non-Recipe Items Reported in the Final List

	Mean
Standard Intro.	13.1
(s.d.)	(3.4)
Think Aloud Intro.	14.0
(s.d.)	(4.6)

$T=-0.50 \quad p>|T|=0.63$

The mean number of food and beverage items for the two groups are now quite similar. The standard group has a mean of 13.1 foods and beverages, and the cognitive group has a mean of 14.0, which is not a statistically significant difference ($p=0.63$). This suggest that if the numerical differences observed for the Final List in Table 2 were indicative of true differences, then a possible explanation could be the disparity in the number of home-made foods reported by respondents of each group.

Given the small number of cases in this study, the findings should be considered tentative. With only ten cases in each group, it is unlikely that differences will be statistically significant. However the initial results suggest that the Quick List question, when administered as a standard interview, does as good a job of eliciting reports of foods and beverages as a think aloud interview. This may be a result of the specific wording of the question. From observing the interviews it seems that both groups reported in a think aloud style. The standard group was never given any instructions to do so, but as mentioned previously, the wording of the question seemed to lead the respondent to use two cognitive strategies often encountered in think aloud interviews: chronological reporting of events, and retrieval by association.

In terms of the sponsors for the CSFII instrument, this wording is optimal. On the other hand, this question may not be the best for evaluating the applicability of cognitive research results in a field setting.

Conclusions

A review of the literature suggests that the thinking aloud procedure may affect recall. Work by several authors suggests that certain characteristics of a concurrent think aloud interview with probing may actually improve recall: emphasizing the response process rather than the actual answer, a slower paced interview, and increased saliency of some information. Unfortunately, these results were not replicated here. This could be due to, as noted above, the wording of the test question itself. In any case, given the small sample size it is difficult to say with any certainty what the reasons may be for our differing results. Additional work in the area should include at least 43 interviews per treatment group in order to detect real differences of the magnitude measured in this study (alpha=.10).

It is fast becoming a standard practice to conduct cognitive testing of questionnaires which will be administered as standard interviews in the field, though there is little research available which addresses the possible differences in response that may result from each method of interviewing. As this method of pretesting questionnaires becomes more common, it is necessary to continue research in this area.

REFERENCES

1. Burton, S. & Blair, E. (1991) "Task Conditions, Response Formulation Processes, and Response Accuracy for Behavioral Frequency Questions in Surveys" Public Opinion Quarterly, 55.

2. Cannell, C.F., Miller, P.V., & Oskenberg, L. (1981) "Research on Interviewing Techniques" in S. Leinhardt (Ed.) Sociological Methodolgy, San Francisco, Jossey-Bass.

3. DeMaio, T., Ciochetto, S. & Davis, W. (1993) "Research on the Continuing Survey of Food Intakes by Individuals" Paper prepared for the Annual Meetings of the American Association for Public Opinion Research, St. Charles, IL.

4. DeMaio, T., Mathiowetz, N., Rothgeb, J., Beach, M. E., and Durant, S. (1993) "Protocol for Pretesting Demographic Surveys" Unpublished Census Bureau manuscript.

5. Ericsson, K. A. & Simon, H. A. (1980) "Verbal Reports as Data" Psychological Review, 87, 3.

6. Jobe, J. B. & Mingay, D. J. (1990) "Cognitive Laboratory Approach to Designing Questionnaires for Surveys of the Elderly" Public Health Reports, 105, 5.

7. Jobe, J. B. & Mingay, D. J. (1989) "Cognitive Research Improves Questionnaires" American Journal of Public Health, 79, 8.

8. Kahneman, D., & Tversky, A., (1971) "Subjective Probability: A Judgement of Representative-ness," Cognitive Psychology, 3.

9. Krosnick, J.A., & Alwin, D.F., (1987) "An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement," Public Opinion Quarterly, 51, 2.

10. Laurent, A. Cannell, C. F. & Marquis, K. H. (1972) "Reporting Health Events in Household Interviews: Effects of an Extensive Questionnaire and Diary Procedure" Vital and Health Statistics, 2, 49, Washington DC: U.S. Government Printing Office.

11. Means, B., Swan, G. E., Jobe, J. B., Esposito, J. L. & Loftus, E. F. (1989) "Recall Strategies for Estimation of Smoking Levels in Health Surveys" Paper for the American Statistical Association Meetings, Washington DC.

12. Means, B., Swan, G.E., Jobe, J.B. & Esposito, J.L. (1991) "An Alternative Approach to Obtaining Personal History Data" In P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz & S.Sudman (Eds.) Measurement Errors in Surveys New York: Wiley.

13. Means, B. & Loftus, E.F. (1991) "When Personal History Repeats Itself: Decomposing Memories for Recurring Events" Applied Cognitive Psychology, 5.

14. Mingay, D. J. & Greenwell, M. T. (1989) "Memory Bias and Response-Order Effects" Journal of Official Statistics, 5

15. Tulving, E. (1983) Elements of Episodic Memory Oxford: Clarendon Press.

16. Willis, G. B., Royston, P. & Bercini, D. (1991) "The Use of Verbal Report Methods in the Development and Testing of Survey Questionnaires" Applied Cognitive Psychology, 5.

NOTES

1. For a review of cognitive interviews as a questionnaire pretesting methodology see DeMaio et al. (1993).